Doing Science Faithfully

Or: can a niche application to evaluate charges of cheating at chess inform statistical practice in medicine, both for focus on human reality and defending against drives to unreality?

Kenneth W. Regan[1]
University at Buffalo (SUNY)

17 April, 2026

[1]With grateful acknowledgment to co-authors Guy L. Haworth and Tamal Biswas, students in my graduate seminars, and UB's Center for Computational Research (CCR)

Some personal and professional background

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986. Postdocs at Oxford and Cornell.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986. Postdocs at Oxford and Cornell.
- UB August 1989, tenure in 1995, full professor 2022.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986. Postdocs at Oxford and Cornell.
- UB August 1989, tenure in 1995, full professor 2022.
- Combinatorial Mathematics, then Computational Complexity.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986. Postdocs at Oxford and Cornell.
- UB August 1989, tenure in 1995, full professor 2022.
- Combinatorial Mathematics, then Computational Complexity.
- Quantum Computing neighbors complexity.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986. Postdocs at Oxford and Cornell.
- UB August 1989, tenure in 1995, full professor 2022.
- Combinatorial Mathematics, then Computational Complexity.
- Quantum Computing neighbors complexity. **Shor's Theorem**.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986. Postdocs at Oxford and Cornell.
- UB August 1989, tenure in 1995, full professor 2022.
- Combinatorial Mathematics, then Computational Complexity.
- Quantum Computing neighbors complexity. **Shor's Theorem**.
- Co-wrote textbook with Richard Lipton.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986. Postdocs at Oxford and Cornell.
- UB August 1989, tenure in 1995, full professor 2022.
- Combinatorial Mathematics, then Computational Complexity.
- Quantum Computing neighbors complexity. **Shor's Theorem**.
- Co-wrote textbook with Richard Lipton.
- Also co-wrote his weblog Gödel's Lost Letter and P=NP, 2009–2024.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986. Postdocs at Oxford and Cornell.
- UB August 1989, tenure in 1995, full professor 2022.
- Combinatorial Mathematics, then Computational Complexity.
- Quantum Computing neighbors complexity. **Shor's Theorem**.
- Co-wrote textbook with Richard Lipton.
- Also co-wrote his weblog Gödel's Lost Letter and P=NP, 2009–2024.
- All the while, I resisted entreaties to do computer chess.

Some personal and professional background

- Grew up in Paramus, NJ, became a USCF Master at age 13.
- US Junior Champion (first equal), 1977.
- FIDE International Master title, 1981. (I am not a Grandmaster.)
- BA (not BS) in Mathematics, Princeton University, 1981.
- Oxford D.Phil. (not PhD), 1986. Postdocs at Oxford and Cornell.
- UB August 1989, tenure in 1995, full professor 2022.
- Combinatorial Mathematics, then Computational Complexity.
- Quantum Computing neighbors complexity. **Shor's Theorem**.
- Co-wrote textbook with Richard Lipton.
- Also co-wrote his weblog Gödel's Lost Letter and P=NP, 2009–2024.
- All the while, I resisted entreaties to do computer chess.
- Then came the 2006 World Championship Cheating Allegation.

About the title...

About the title...

- Google “do science faithfully” in quotes.

About the title...

- Google “do science faithfully” in quotes. ISCAST (2022 interview)

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.
- My 2007 webpage—earliest use?

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.
- My 2007 webpage—earliest use? Also this, this.

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.
- My 2007 webpage—earliest use? Also this, this.
- Older Google N-grams seem to find only "con- science faithfully."

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.
- My 2007 webpage—earliest use? Also this, this.
- Older Google N-grams seem to find only "con- science faithfully."
- I did intend religious reference—I'd been dialoguing with Frederic Friedel of Chessbase, about whom here.

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.
- My 2007 webpage—earliest use? Also this, this.
- Older Google N-grams seem to find only "con- science faithfully."
- I did intend religious reference—I'd been dialoguing with Frederic Friedel of Chessbase, about whom here. Friedel told me about Richard Dawkins and about his own visits to the Center for Skeptical Inquiry near UB (before my time).

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.
- My 2007 webpage—earliest use? Also this, this.
- Older Google N-grams seem to find only "con- science faithfully."
- I did intend religious reference—I'd been dialoguing with Frederic Friedel of Chessbase, about whom here. Friedel told me about Richard Dawkins and about his own visits to the Center for Skeptical Inquiry near UB (before my time).
- It meant my rejecting Stephen Jay Gould's NOMA.

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.
- My 2007 webpage—earliest use? Also this, this.
- Older Google N-grams seem to find only "con- science faithfully."
- I did intend religious reference—I'd been dialoguing with Frederic Friedel of Chessbase, about whom here. Friedel told me about Richard Dawkins and about his own visits to the Center for Skeptical Inquiry near UB (before my time).
- It meant my rejecting Stephen Jay Gould's NOMA.
- Not about "facts" separate from "values": statistics mediates **grounds for assent**.

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.
- My 2007 webpage—earliest use? Also this, this.
- Older Google N-grams seem to find only "con- science faithfully."
- I did intend religious reference—I'd been dialoguing with Frederic Friedel of Chessbase, about whom here. Friedel told me about Richard Dawkins and about his own visits to the Center for Skeptical Inquiry near UB (before my time).
- It meant my rejecting Stephen Jay Gould's NOMA.
- Not about "facts" separate from "values": statistics mediates **grounds for assent**.
- Phrase can be read completely secularly: faithfulness to science.

About the title...

- Google "do science faithfully" in quotes. ISCAST (2022 interview)
- Google's "AI Overview" is quite good.
- My 2007 webpage—earliest use? Also this, this.
- Older Google N-grams seem to find only "con- science faithfully."
- I did intend religious reference—I'd been dialoguing with Frederic Friedel of Chessbase, about whom here. Friedel told me about Richard Dawkins and about his own visits to the Center for Skeptical Inquiry near UB (before my time).
- It meant my rejecting Stephen Jay Gould's NOMA.
- Not about "facts" separate from "values": statistics mediates **grounds for assent**.
- Phrase can be read completely secularly: faithfulness to science.
- But also: pursuit of *transgression* needs to allow for *redemption*.

Doing Science Recklessly...

Doing Science Recklessly...

- Newest major chess-playing program Reckless.

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0...

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)
- These programs **slay** us now, even on smartphones.

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)
- These programs **slay** us now, even on smartphones.
- FIDE Candidates 2026 just ended.

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)
- These programs **slay** us now, even on smartphones.
- FIDE Candidates 2026 just ended. Blitzkrieg by dark horse Javokhir Sindarov of Uzbekistan.

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)
- These programs **slay** us now, even on smartphones.
- FIDE Candidates 2026 just ended. Blitzkrieg by dark horse Javokhir Sindarov of Uzbekistan.
- Was remarked during key round-5 victory that JS was matching every move by SF18.

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)
- These programs **slay** us now, even on smartphones.
- FIDE Candidates 2026 just ended. Blitzkrieg by dark horse Javokhir Sindarov of Uzbekistan.
- Was remarked during key round-5 victory that JS was matching every move by SF18.
- I reproduce this. With Reckless too.

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)
- These programs **slay** us now, even on smartphones.
- FIDE Candidates 2026 just ended. Blitzkrieg by dark horse Javokhir Sindarov of Uzbekistan.
- Was remarked during key round-5 victory that JS was matching every move by SF18.
- I reproduce this. With Reckless too.
- Proof of cheating? Many have regarded it so...

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)
- These programs **slay** us now, even on smartphones.
- FIDE Candidates 2026 just ended. Blitzkrieg by dark horse Javokhir Sindarov of Uzbekistan.
- Was remarked during key round-5 victory that JS was matching every move by SF18.
- I reproduce this. With Reckless too.
- Proof of cheating? Many have regarded it so... Survey

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)
- These programs **slay** us now, even on smartphones.
- FIDE Candidates 2026 just ended. Blitzkrieg by dark horse Javokhir Sindarov of Uzbekistan.
- Was remarked during key round-5 victory that JS was matching every move by SF18.
- I reproduce this. With Reckless too.
- Proof of cheating? Many have regarded it so... Survey
- Even highly responsible work can overlook some external things.

Doing Science Recklessly...

- Newest major chess-playing program Reckless.
- Beat perennial #2 Leela Chess Zero but lost to Stockfish 18.
- Is easier to mount than LC0... hence easier to cheat with(?)
- These programs **slay** us now, even on smartphones.
- FIDE Candidates 2026 just ended. Blitzkrieg by dark horse Javokhir Sindarov of Uzbekistan.
- Was remarked during key round-5 victory that JS was matching every move by SF18.
- I reproduce this. With Reckless too.
- Proof of cheating? Many have regarded it so... Survey
- Even highly responsible work can overlook some external things.
- I have not fully calibrated either Reckless or SF18 yet.

Analytics Versus **Predictive** Analytics

To have a Predictive Analytic Model **IMPHO** means that it:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$

Analytics Versus **Predictive** Analytics

To have a Predictive Analytic Model **IMPHO** means that it:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.

Analytics Versus **Predictive** Analytics

To have a Predictive Analytic Model **IMPHO** means that it:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.

Analytics Versus **Predictive** Analytics

To have a Predictive Analytic Model **IMPHO** means that it:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

Analytics Versus Predictive Analytics

To have a Predictive Analytic Model **IMPHO** means that it:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

In a **utility-based** model, each m_i has a utility or cost u_i.
Main risk/reward quantity then becomes $E = \sum_i p_i u_i$.

Analytics Versus **Predictive** Analytics

To have a Predictive Analytic Model **IMPHO** means that it:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

In a **utility-based** model, each m_i has a utility or cost u_i.
Main risk/reward quantity then becomes $E = \sum_i p_i u_i$.

- **Insurance**: m_i are risk factors; costs u_i need not influence p_i.

Analytics Versus Predictive Analytics

To have a Predictive Analytic Model **IMPHO** means that it:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

In a **utility-based** model, each m_i has a utility or cost u_i.
Main risk/reward quantity then becomes $E = \sum_i p_i u_i$.

- **Insurance**: m_i are risk factors; costs u_i need not influence p_i.
- **Chess**: m_i are legal moves; u_i are engine values and influence p_i.

Analytics Versus **Predictive** Analytics

To have a Predictive Analytic Model **IMPHO** means that it:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \dots, m_j, \dots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

In a **utility-based** model, each m_i has a utility or cost u_i.
Main risk/reward quantity then becomes $E = \sum_i p_i u_i$.

- **Insurance**: m_i are risk factors; costs u_i need not influence p_i.
- **Chess**: m_i are legal moves; u_i are engine values and influence p_i.
- **Multiple-choice tests**: m_i are possible answers to a test question, u_i = gain/loss for right/wrong answer.

Chess and Tests—With Partial Credits (Or LLMs?)

The ____ of drug-resistant strains of bacteria and viruses has ____ researchers' hopes that permanent victories against many diseases have been achieved.

(a) vigor . . corroborated

(b) feebleness . . dashed

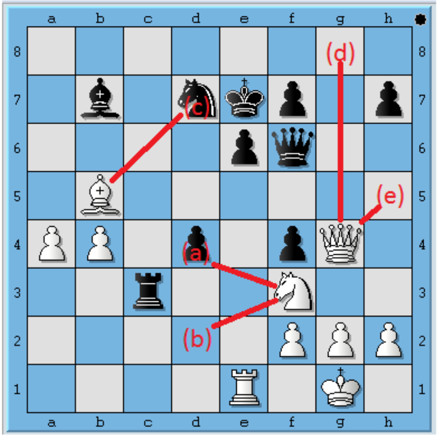
(c) proliferation . . blighted

(d) destruction . . disputed

(e) disappearance . . frustrated

(source: itunes.apple.com)

=



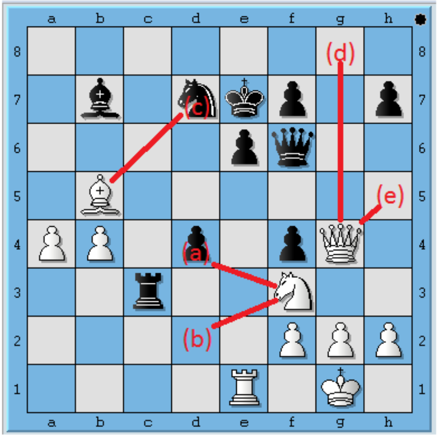
Chess and Tests—With Partial Credits (Or LLMs?)

The ____ of drug-resistant strains of bacteria and viruses has ____ researchers' hopes that permanent victories against many diseases have been achieved.

- (a) vigor . . corroborated
- (b) feebleness . . dashed
- (c) proliferation . . blighted
- (d) destruction . . disputed
- (e) disappearance . . frustrated

(source: itunes.apple.com)

=



Here (b,c) are **equal-optimal** choices, (a) is bad, but (d) and (e) are reasonable—worth part credit.

Chess and Tests—With Partial Credits (Or LLMs?)

The ___ of drug-resistant strains of bacteria and viruses has ___ researchers' hopes that permanent victories against many diseases have been achieved.

(a) vigor . . corroborated

(b) feebleness . . dashed

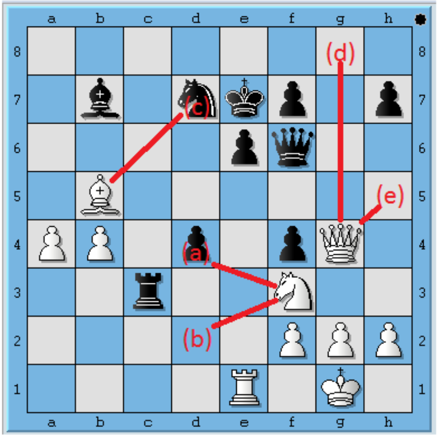
(c) proliferation . . blighted

(d) destruction . . disputed

(e) disappearance . . frustrated

(source: itunes.apple.com)

=



Here (b,c) are **equal-optimal** choices, (a) is bad, but (d) and (e) are reasonable—worth part credit.

Aptitude—Via Elo Grades (calculator)

- Named for **Arpad Elo**, number R_P rates skill of player P.

Aptitude—Via Elo Grades (calculator)

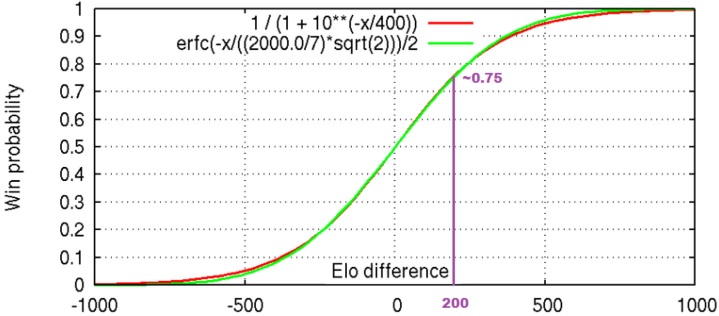
- Named for **Arpad Elo**, number R_P rates skill of player P.
- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.

Aptitude—Via Elo Grades (calculator)

- Named for **Arpad Elo**, number R_P rates skill of player P.
- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**

Aptitude—Via Elo Grades (calculator)

- Named for **Arpad Elo**, number R_P rates skill of player P.
- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**
- Expectation given by rating *difference* via this logistic curve:



How The Model Operates

How The Model Operates

- Take parameters s, c, h from a player's Elo rating R (or skill profile).

How The Model Operates

- Take parameters s, c, h from a player's Elo rating R (or skill profile).
- Generate probability p_i for each legal move m_i.

How The Model Operates

- Take parameters s, c, h from a player's Elo rating R (or skill profile).
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, **1,000**p_i times.

How The Model Operates

- Take parameters s, c, h from a player's Elo rating R (or skill profile).
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, **1,000**p_i times.
- **Roll the die** to give confidence intervals that go with the p_i.

How The Model Operates

- Take parameters s, c, h from a player's Elo rating R (or skill profile).
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, **1,000**p_i times.
- **Roll the die** to give confidence intervals that go with the p_i.
- (Correct after-the-fact for chess decisions not being independent.)

Main Outputs:

How The Model Operates

- Take parameters s, c, h from a player's Elo rating R (or skill profile).
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, **1,000**p_i times.
- **Roll the die** to give confidence intervals that go with the p_i.
- (Correct after-the-fact for chess decisions not being independent.)

Main Outputs:

- **Statistical z-scores** for various $(actual - projected)$ quantities:
  - **T1-match**: Agreement with the move listed first by the computer.
  - **EV-match**: Includes moves of equal-optimal value not listed first.
  - **ASD**: Average *scaled* difference in value from inferior moves.

How The Model Operates

- Take parameters s, c, h from a player's Elo rating R (or skill profile).
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, **1,000**p_i times.
- **Roll the die** to give confidence intervals that go with the p_i.
- (Correct after-the-fact for chess decisions not being independent.)

Main Outputs:

- **Statistical z-scores** for various ($actual - projected$) quantities:
  - **T1-match**: Agreement with the move listed first by the computer.
  - **EV-match**: Includes moves of equal-optimal value not listed first.
  - **ASD**: Average *scaled* difference in value from inferior moves.
- An **Intrinsic Performance Rating** (**IPR**) for the set of games.

Fit s, c, h by making T1,EV,ASD be **unbiased estimators** on the training sets, which are stratified by Elo ratings.

Self-Regulation in Chess and Medicine

Self-Regulation in Chess and Medicine

- How accurately can you bound your own predictive error?

Self-Regulation in Chess and Medicine

- How accurately can you bound your own predictive error?
- Sir David Spiegelhalter created a tool.

Self-Regulation in Chess and Medicine

- How accurately can you bound your own predictive error?
- Sir David Spiegelhalter created a tool.
- Aimed at medical diagnostic accuracy, but I use it for chess.

Self-Regulation in Chess and Medicine

- How accurately can you bound your own predictive error?
- Sir David Spiegelhalter created a tool.
- Aimed at medical diagnostic accuracy, but I use it for chess.
- Shows that relative log-error is mostly within 5%.

Self-Regulation in Chess and Medicine

- How accurately can you bound your own predictive error?
- Sir David Spiegelhalter created a tool.
- Aimed at medical diagnostic accuracy, but I use it for chess.
- Shows that relative log-error is mostly within 5%.
- Armed with that reassurance, I then turn this z-of-Brier score into another test of player deviance.

Self-Regulation in Chess and Medicine

- How accurately can you bound your own predictive error?
- Sir David Spiegelhalter created a tool.
- Aimed at medical diagnostic accuracy, but I use it for chess.
- Shows that relative log-error is mostly within 5%.
- Armed with that reassurance, I then turn this z-of-Brier score into another test of player deviance.
- Analogy: if you know your rain-forecasting accuracy, but 100-year floods happen every 5 years, you may conclude that Mother Nature is cheating with a systematic heat source.

Self-Regulation in Chess and Medicine

- How accurately can you bound your own predictive error?
- Sir David Spiegelhalter created a tool.
- Aimed at medical diagnostic accuracy, but I use it for chess.
- Shows that relative log-error is mostly within 5%.
- Armed with that reassurance, I then turn this z-of-Brier score into another test of player deviance.
- Analogy: if you know your rain-forecasting accuracy, but 100-year floods happen every 5 years, you may conclude that Mother Nature is cheating with a systematic heat source.
- This test is so far no stronger or weaker than my original based on MM and CPL metrics.

What People Desire From Tools

In my January 2007 phone call with Frederic Friedel, we discussed what chess officials would desire from a tool. This quote stood out:

> “People want to be able to push a button, and get a number.”

I organized my system to do exactly the opposite:

What People Desire From Tools

In my January 2007 phone call with Frederic Friedel, we discussed what chess officials would desire from a tool. This quote stood out:

> "People want to be able to push a button, and get a number."

I organized my system to do exactly the opposite:

- It has an automated **screening** stage that is not for judgment.

What People Desire From Tools

In my January 2007 phone call with Frederic Friedel, we discussed what chess officials would desire from a tool. This quote stood out:

> "People want to be able to push a button, and get a number."

I organized my system to do exactly the opposite:

- It has an automated **screening** stage that is not for judgment. It triages players and gives ooverall big-picture positive feedback on the health of a competition.

What People Desire From Tools

In my January 2007 phone call with Frederic Friedel, we discussed what chess officials would desire from a tool. This quote stood out:

> "People want to be able to push a button, and get a number."

I organized my system to do exactly the opposite:

- It has an automated **screening** stage that is not for judgment. It triages players and gives ooverall big-picture positive feedback on the health of a competition.
- Its **full test** stage must be manually operated.

What People Desire From Tools

In my January 2007 phone call with Frederic Friedel, we discussed what chess officials would desire from a tool. This quote stood out:

> "People want to be able to push a button, and get a number."

I organized my system to do exactly the opposite:

- It has an automated **screening** stage that is not for judgment. It triages players and gives ooverall big-picture positive feedback on the health of a competition.
- Its **full test** stage must be manually operated.
- Prime directive is to avoid a "rush to judgment." Example.

What People Desire From Tools

In my January 2007 phone call with Frederic Friedel, we discussed what chess officials would desire from a tool. This quote stood out:

"People want to be able to push a button, and get a number."

I organized my system to do exactly the opposite:

- It has an automated **screening** stage that is not for judgment. It triages players and gives ooverall big-picture positive feedback on the health of a competition.
- Its **full test** stage must be manually operated.
- Prime directive is to avoid a "rush to judgment." Example.
- In general, my main AI fear is not human extinction

What People Desire From Tools

In my January 2007 phone call with Frederic Friedel, we discussed what chess officials would desire from a tool. This quote stood out:

> "People want to be able to push a button, and get a number."

I organized my system to do exactly the opposite:

- It has an automated **screening** stage that is not for judgment. It triages players and gives ooverall big-picture positive feedback on the health of a competition.
- Its **full test** stage must be manually operated.
- Prime directive is to avoid a "rush to judgment." Example.
- In general, my main AI fear is not human extinction—but rather that we will come to outsource our hard decisions to AI.

Z-Scores

- A **z-score** measures performance relative to natural expectation.

Z-Scores

- A **z-score** measures performance relative to natural expectation.
- Used extensively by business in Quality Assurance, Human Resources Management, and by many testing agencies.

Z-Scores

- A **z-score** measures performance relative to natural expectation.
- Used extensively by business in Quality Assurance, Human Resources Management, and by many testing agencies.
- Expressed in units of standard deviations, called "sigmas" (σ).

Z-Scores

- A **z-score** measures performance relative to natural expectation.
- Used extensively by business in Quality Assurance, Human Resources Management, and by many testing agencies.
- Expressed in units of standard deviations, called “sigmas” (σ).
- Correspond to statements of odds-against (**but see next slides**):

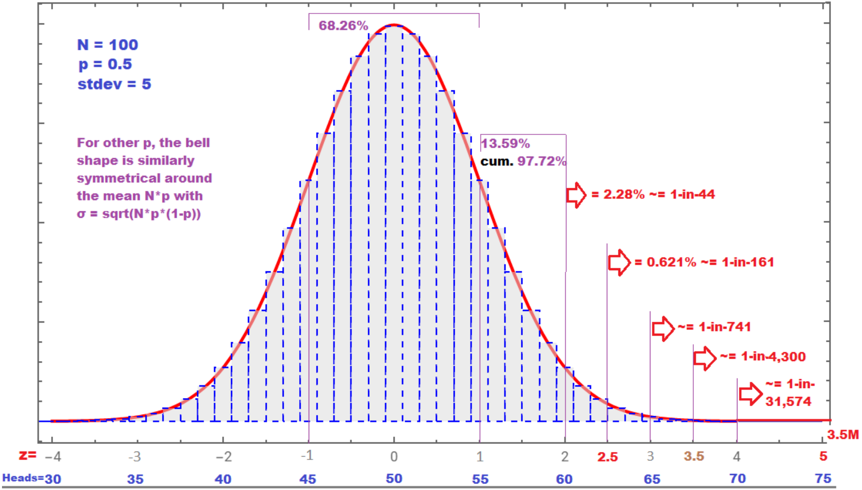
Z-Scores

- A **z-score** measures performance relative to natural expectation.
- Used extensively by business in Quality Assurance, Human Resources Management, and by many testing agencies.
- Expressed in units of standard deviations, called “sigmas” (σ).
- Correspond to statements of odds-against (**but see next slides**):
- “Six Sigma” (6σ) means about 500,000,000–1 odds;
- 5σ = 3,000,000–1;
- 4.75σ = 1,000,000–1;
- 4.5σ = 300,000–1;

Z-Scores

- A **z-score** measures performance relative to natural expectation.
- Used extensively by business in Quality Assurance, Human Resources Management, and by many testing agencies.
- Expressed in units of standard deviations, called "sigmas" (σ).
- Correspond to statements of odds-against (**but see next slides**):
- "Six Sigma" (6σ) means about 500,000,000–1 odds;
- 5σ = 3,000,000–1;
- 4.75σ = 1,000,000–1;
- 4.5σ = 300,000–1;
- 4σ = 32,000–1;
- 3σ = 740–1;
- 2σ = 43–1 (civil minimum standard, polling "margin of error").

Bell Curve and Tails



Blue = binomial 100 scale of the **screening stage**. WSTC examples.

Suppose We Get $z = 3.54$

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000.

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives “raw face-value odds” of “5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.
  - 1-in-50 (greater for kids) to 1-in-200 for online chess.

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.
  - 1-in-50 (greater for kids) to 1-in-200 for online chess.
- **Look-Elsewhere Effect**: How many were playing chess that day?

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
    - 1-in-5,000 to 1-in-10,000 for in-person chess.
    - 1-in-50 (greater for kids) to 1-in-200 for online chess.
- **Look-Elsewhere Effect**: How many were playing chess that day? weekend?

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.
  - 1-in-50 (greater for kids) to 1-in-200 for online chess.
- **Look-Elsewhere Effect**: How many were playing chess that day? weekend? week?

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.
  - 1-in-50 (greater for kids) to 1-in-200 for online chess.
- **Look-Elsewhere Effect**: How many were playing chess that day? weekend? week? month?

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.
  - 1-in-50 (greater for kids) to 1-in-200 for online chess.
- **Look-Elsewhere Effect**: How many were playing chess that day? weekend? week? month? year?

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.
  - 1-in-50 (greater for kids) to 1-in-200 for online chess.
- **Look-Elsewhere Effect**: How many were playing chess that day? weekend? week? month? year?
- **"Shiny Marbles Get Noticed"**

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.
  - 1-in-50 (greater for kids) to 1-in-200 for online chess.
- **Look-Elsewhere Effect**: How many were playing chess that day? weekend? week? month? year?
- **"Shiny Marbles Get Noticed"**—and this influences the **conditional probability** associated to a possibly suspicious observation.

Over large datasets from (presumably) non-cheating players, the **Central Limit Theorem** "kicks in" well: the z-scores conform to the bell curve. Example Spreadsheets.

Some Example Cases (old ones on-purpose...)

Cheating and ...

- Sebastien Feller, 2010 Olympiad, rated **2649**.
  - 4 confessed all-cheating games: **z=2.96** with IPR **3240**.
  - 5 other games: IPR **2547**.
  - Fact of on-site evidence made these results significant.
- Borislav Ivanov, 2012 Zadar Open, rating 2227→2342.
  - Z-scores as high as **5.10**.
  - IPR near **3100**.
  - FIDE now allows verdict "assumed cheating" by stats alone.

[Results from model built using old Rybka 3 engine]

Non-Cheating

- Kramnik-Topalov World Championship Match, 2006
  - Topalov's manager accused Kramnik's moves in games 1—6 with the engine Fritz 9.
  - I reproduced the claimed 90% concordance only in the second half of Game 2.
  - Still matches 26-of-32 (**81%**) to both Stockfish 11 & 16.
  - But my model projects **82%** concordance there---most of those moves were "forced" hence relatively easy to find.

Human Studies and Chess Research

Especially in social science research, many results come from studies that

1. are well-targeted to the concept and hypothesis, but

Human Studies and Chess Research

Especially in social science research, many results come from studies that

1. are well-targeted to the concept and hypothesis, but
2. have under 100 test subjects...

Human Studies and Chess Research

Especially in social science research, many results come from studies that

1. are well-targeted to the concept and hypothesis, but
2. have under 100 test subjects...
3. ...under simulated conditions...

Human Studies and Chess Research

Especially in social science research, many results come from studies that

1. are well-targeted to the concept and hypothesis, but
2. have under 100 test subjects...
3. ...under simulated conditions...
4. ...with unclear metrics and alignment of personal vs. test goals...,

Human Studies and Chess Research

Especially in social science research, many results come from studies that

1. are well-targeted to the concept and hypothesis, but
2. have under 100 test subjects...
3. ...under simulated conditions...
4. ...with unclear metrics and alignment of personal vs. test goals...,
5. ...and where reproducibility is doubtful and arduous.

Per my Daniel Kahneman obit, we should trade 1 against wealth of 2,3,4,5: lots of players and games, real competition, clear goals and metrics, reproducible, and conducive to abundant falsifiable predictions.

Human Studies and Chess Research

Especially in social science research, many results come from studies that

1. are well-targeted to the concept and hypothesis, but
2. have under 100 test subjects...
3. ...under simulated conditions...
4. ...with unclear metrics and alignment of personal vs. test goals...,
5. ...and where reproducibility is doubtful and arduous.

Per my Daniel Kahneman obit, we should trade 1 against wealth of 2,3,4,5: lots of players and games, real competition, clear goals and metrics, reproducible, and conducive to abundant falsifiable predictions.

My model affords multiple ways of **cross-validation**.

Human Studies and Chess Research

Especially in social science research, many results come from studies that

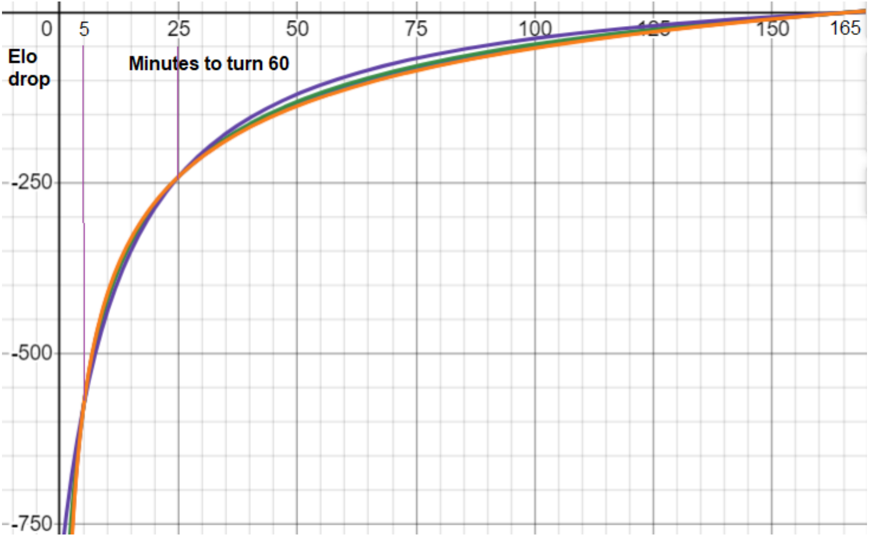
1. are well-targeted to the concept and hypothesis, but
2. have under 100 test subjects...
3. ...under simulated conditions...
4. ...with unclear metrics and alignment of personal vs. test goals...,
5. ...and where reproducibility is doubtful and arduous.

Per my Daniel Kahneman obit, we should trade 1 against wealth of 2,3,4,5: lots of players and games, real competition, clear goals and metrics, reproducible, and conducive to abundant falsifiable predictions.

My model affords multiple ways of **cross-validation**.

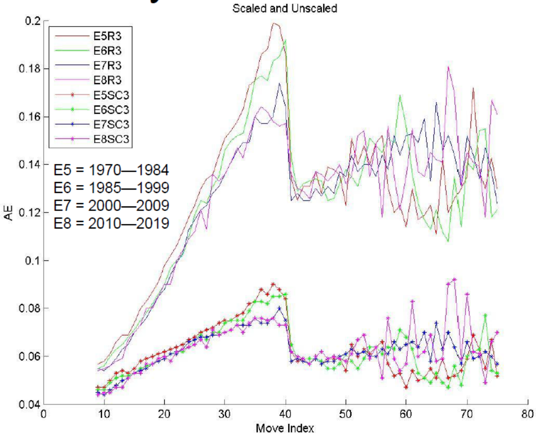
Let's consider elements of **difficulty** and **time usage**.

Time-Quality Curves (whole graph)



Time Usage, Procrastination, and Centipawn Loss

Error By Move Number in Games



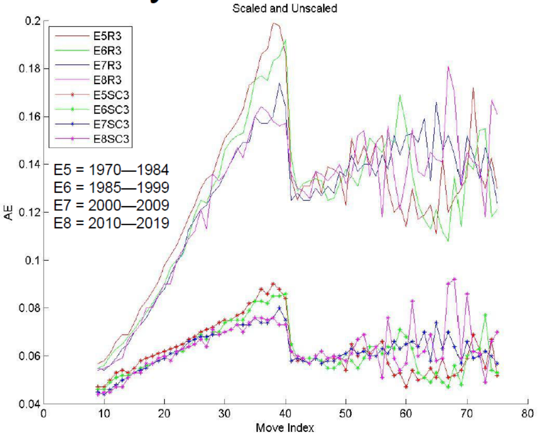
Effect of time pressure approaching Move 40 is clear.

Moves 17—32 bridge between opening theory and worst of Zeitnot.

Mainly tournaments with lump of extra time after turn 40 up thru 2015.

Time Usage, Procrastination, and Centipawn Loss

Error By Move Number in Games



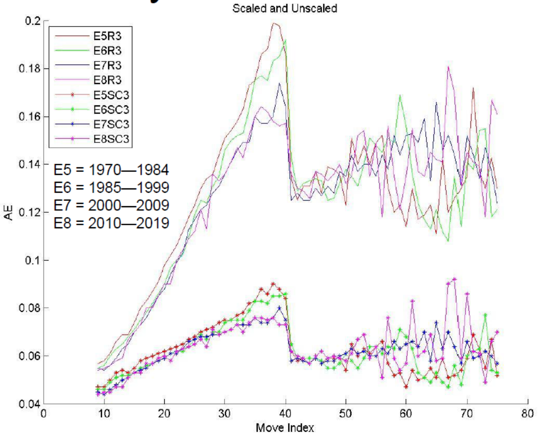
Effect of time pressure approaching Move 40 is clear.

Moves 17—32 bridge between opening theory and worst of Zeitnot.

Mainly tournaments with lump of extra time after turn 40 up thru 2015. Can imagine worse curve without a turn-40 sum (even with increment).

Time Usage, Procrastination, and Centipawn Loss

Error By Move Number in Games



Effect of time pressure approach ing Move 40 is clear.

Moves 17—32 bridge between opening theory and worst of Zeitnot.

Mainly tournaments with lump of extra time after turn 40 up thru 2015.
Can imagine worse curve without a turn-40 sum (even with increment).
(How) Can we teach players to use time with more foresight?

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

“Thinking Is Bad For You”

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move:

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.
- At most **5** seconds (sample size 605): **3230 +- 160**.

How about moves with long thinks—?

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.
- At most **5** seconds (sample size 605): **3230 +- 160**.

How about moves with long thinks—?

- Positions with 5–10 minutes consumed:

“Thinking Is Bad For You”

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.
- At most **5** seconds (sample size 605): **3230 +- 160**.

How about moves with long thinks—?

- Positions with 5–10 minutes consumed: **1460 +- 85**.

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.
- At most **5** seconds (sample size 605): **3230 +- 160**.

How about moves with long thinks—?

- Positions with 5–10 minutes consumed: **1460 +- 85**.
- Using 10–15 minutes (705 positions): **1235 +- 170**.

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.
- At most **5** seconds (sample size 605): **3230 +- 160**.

How about moves with long thinks—?

- Positions with 5–10 minutes consumed: **1460 +- 85**.
- Using 10–15 minutes (705 positions): **1235 +- 170**.
- Using $\geq$ 15 minutes (371 positions): **1410 +- 205**.

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.
- At most **5** seconds (sample size 605): **3230 +- 160**.

How about moves with long thinks—?

- Positions with 5–10 minutes consumed: **1460 +- 85**.
- Using 10–15 minutes (705 positions): **1235 +- 170**.
- Using $\geq$ 15 minutes (371 positions): **1410 +- 205**.
- Vivid reproduction of [SZS 2022] (and also Anderson et al., 2016 thru now for online blitz).

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.
- At most **5** seconds (sample size 605): **3230 +- 160**.

How about moves with long thinks—?

- Positions with 5–10 minutes consumed: **1460 +- 85**.
- Using 10–15 minutes (705 positions): **1235 +- 170**.
- Using $\geq$ 15 minutes (371 positions): **1410 +- 205**.
- Vivid reproduction of [SZS 2022] (and also Anderson et al., 2016 thru now for online blitz). "Think before you act...

"Thinking Is Bad For You"

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.
- At most **5** seconds (sample size 605): **3230 +- 160**.

How about moves with long thinks—?

- Positions with 5–10 minutes consumed: **1460 +- 85**.
- Using 10–15 minutes (705 positions): **1235 +- 170**.
- Using $\geq$ 15 minutes (371 positions): **1410 +- 205**.
- Vivid reproduction of [SZS 2022] (and also Anderson et al., 2016 thru now for online blitz). "Think before you act...but not too long."

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- Positions on which they spent at most **30 seconds** on the move:

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- Positions on which they spent at most **30 seconds** on the move: **2170 +- 125** (2,996 pos.)

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- Positions on which they spent at most **30 seconds** on the move: **2170 +- 125** (2,996 pos.)
- At most **10** seconds: **2860 +- 245** (632 positions)

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- Positions on which they spent at most **30 seconds** on the move: **2170 +- 125** (2,996 pos.)
- At most **10** seconds: **2860 +- 245** (632 positions)
- At most **5** seconds (sample size 151): **2935 +- 555**.

How about when little kids think longer?

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- Positions on which they spent at most **30 seconds** on the move: **2170 +- 125** (2,996 pos.)
- At most **10** seconds: **2860 +- 245** (632 positions)
- At most **5** seconds (sample size 151): **2935 +- 555**.

How about when little kids think longer?

- Positions with 5–10 minutes consumed (729 pos.):

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- Positions on which they spent at most **30 seconds** on the move: **2170 +- 125** (2,996 pos.)
- At most **10** seconds: **2860 +- 245** (632 positions)
- At most **5** seconds (sample size 151): **2935 +- 555**.

How about when little kids think longer?

- Positions with 5–10 minutes consumed (729 pos.): **650 +- 235**.

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- Positions on which they spent at most **30 seconds** on the move: **2170 +- 125** (2,996 pos.)
- At most **10** seconds: **2860 +- 245** (632 positions)
- At most **5** seconds (sample size 151): **2935 +- 555**.

How about when little kids think longer?

- Positions with 5–10 minutes consumed (729 pos.): **650 +- 235**.
- Using 10–15 minutes (168 positions): **465 +- 565**.

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- Positions on which they spent at most **30 seconds** on the move: **2170 +- 125** (2,996 pos.)
- At most **10** seconds: **2860 +- 245** (632 positions)
- At most **5** seconds (sample size 151): **2935 +- 555**.

How about when little kids think longer?

- Positions with 5–10 minutes consumed (729 pos.): **650 +- 235**.
- Using 10–15 minutes (168 positions): **465 +- 565**.
- Using $\geq$ 15 minutes (104 positions): **700 +- 505**.

Thinking is Bad For 8-Year-Olds Too!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating **1596**.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- Positions on which they spent at most **30 seconds** on the move: **2170 +- 125** (2,996 pos.)
- At most **10** seconds: **2860 +- 245** (632 positions)
- At most **5** seconds (sample size 151): **2935 +- 555**.

How about when little kids think longer?

- Positions with 5–10 minutes consumed (729 pos.): **650 +- 235**.
- Using 10–15 minutes (168 positions): **465 +- 565**.
- Using $\geq$ 15 minutes (104 positions): **700 +- 505**.
- **What's going on here?**

Conditional Probability Is Good For You

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.
- Longer than the time to double-check a previously resolved intent.

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.
- Longer than the time to double-check a previously resolved intent.
- **Conditioned** on this, expectations should change.

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.
- Longer than the time to double-check a previously resolved intent.
- **Conditioned** on this, expectations should change.
- When we condition on thinking time left, results are “more normal”:

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.
- Longer than the time to double-check a previously resolved intent.
- **Conditioned** on this, expectations should change.
- When we condition on thinking time left, results are “more normal”:
  - When player has $\leq$ 180 seconds left (633 turns): **1540 +- 280**.

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.
- Longer than the time to double-check a previously resolved intent.
- **Conditioned** on this, expectations should change.
- When we condition on thinking time left, results are “more normal”:
  - When player has $\leq$ 180 seconds left (633 turns): **1540 +- 280**.
  - Or average $\leq$ 60 seconds left to turn 40, not counting increment time: **1685 +- 200**.

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.
- Longer than the time to double-check a previously resolved intent.
- **Conditioned** on this, expectations should change.
- When we condition on thinking time left, results are "more normal":
  - When player has $\leq$ 180 seconds left (633 turns): **1540 +- 280**.
  - Or average $\leq$ 60 seconds left to turn 40, not counting increment time: **1685 +- 200**.
  - Or average 30 seconds left to turn 40, counting half the increment time: **1395 +- 425**.

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.
- Longer than the time to double-check a previously resolved intent.
- **Conditioned** on this, expectations should change.
- When we condition on thinking time left, results are "more normal":
  - When player has ≤ 180 seconds left (633 turns): **1540 +- 280**.
  - Or average ≤ 60 seconds left to turn 40, not counting increment time: **1685 +- 200**.
  - Or average 30 seconds left to turn 40, counting half the increment time: **1395 +- 425**.
- Results still a challenge to Kahneman's "System 2" hypotheses.

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.
- Longer than the time to double-check a previously resolved intent.
- **Conditioned** on this, expectations should change.
- When we condition on thinking time left, results are “more normal”:
  - When player has $\leq$ 180 seconds left (633 turns): **1540 +- 280**.
  - Or average $\leq$ 60 seconds left to turn 40, not counting increment time: **1685 +- 200**.
  - Or average 30 seconds left to turn 40, counting half the increment time: **1395 +- 425**.
- Results still a challenge to Kahneman’s “System 2” hypotheses.
- IMPHO, these results align with the 2007 National Geographic documentary “My Brilliant Brain” with Susan Polgar (crux here).

Conditional Probability Is Good For You

- The fact of thinking long indicates being in a quandary.
- Longer than the time to double-check a previously resolved intent.
- **Conditioned** on this, expectations should change.
- When we condition on thinking time left, results are “more normal”:
  - When player has $\leq$ 180 seconds left (633 turns): **1540 +- 280**.
  - Or average $\leq$ 60 seconds left to turn 40, not counting increment time: **1685 +- 200**.
  - Or average 30 seconds left to turn 40, counting half the increment time: **1395 +- 425**.
- Results still a challenge to Kahneman’s “System 2” hypotheses.
- IMPHO, these results align with the 2007 National Geographic documentary “My Brilliant Brain” with Susan Polgar (crux here).
- **Temptation**: set lower rating R on longer-think moves.

Some General Takeaways

Some General Takeaways

- **Context Matters**.

Some General Takeaways

- **Context Matters**.
- **That we are dealing with people matters**.

Some General Takeaways

- **Context Matters**.
- **That we are dealing with people matters**.
- **“Bang the Metaspace”**—which may be the “meatspace” (as Scott Aaronson calls it).

Some General Takeaways

- **Context Matters**.
- **That we are dealing with people matters**.
- **“Bang the Metaspace”**—which may be the “meatspace” (as Scott Aaronson calls it). AI can especially help here.

Some General Takeaways

- **Context Matters**.
- **That we are dealing with people matters**.
- **"Bang the Metaspace"**—which may be the "meatspace" (as Scott Aaronson calls it). AI can especially help here.
- Many seemingly simply-factual questions have statistical and value qualifiers on their uptake, not just in their wake.

Some General Takeaways

- **Context Matters**.
- **That we are dealing with people matters**.
- **"Bang the Metaspace"**—which may be the "meatspace" (as Scott Aaronson calls it). AI can especially help here.
- Many seemingly simply-factual questions have statistical and value qualifiers on their uptake, not just in their wake.
- **Simple facts are more effective than long studies**.

Some General Takeaways

- **Context Matters**.
- **That we are dealing with people matters**.
- **“Bang the Metaspace”**—which may be the “meatspace” (as Scott Aaronson calls it). AI can especially help here.
- Many seemingly simply-factual questions have statistical and value qualifiers on their uptake, not just in their wake.
- **Simple facts are more effective than long studies**. Especially to answer those who drive toward unreality.

Some General Takeaways

- **Context Matters**.
- **That we are dealing with people matters**.
- **"Bang the Metaspace"**—which may be the "meatspace" (as Scott Aaronson calls it). AI can especially help here.
- Many seemingly simply-factual questions have statistical and value qualifiers on their uptake, not just in their wake.
- **Simple facts are more effective than long studies**. Especially to answer those who drive toward unreality.
- Allowance for people in large cohorts to have "three-sigma down" events.

Some General Takeaways

- **Context Matters**.
- **That we are dealing with people matters**.
- **"Bang the Metaspace"**—which may be the "meatspace" (as Scott Aaronson calls it). AI can especially help here.
- Many seemingly simply-factual questions have statistical and value qualifiers on their uptake, not just in their wake.
- **Simple facts are more effective than long studies**. Especially to answer those who drive toward unreality.
- Allowance for people in large cohorts to have "three-sigma down" events. (But clamp on "four-sigma down" and strive to reduce the variance overall.)

Some General Takeaways

- **Context Matters**.
- **That we are dealing with people matters**.
- **“Bang the Metaspace”**—which may be the “meatspace” (as Scott Aaronson calls it). AI can especially help here.
- Many seemingly simply-factual questions have statistical and value qualifiers on their uptake, not just in their wake.
- **Simple facts are more effective than long studies**. Especially to answer those who drive toward unreality.
- Allowance for people in large cohorts to have “three-sigma down” events. (But clamp on “four-sigma down” and strive to reduce the variance overall.)
- Q & A, Discussion, and Thanks.

Some details on chess programs and model formation

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:
   - As the search progresses, they “change their minds” about which move is best.

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:
    - As the search progresses, they “change their minds” about which move is best.
    - List top move at each depth of search and whether it “matches” the move that was played (MM).

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:
    - As the search progresses, they "change their minds" about which move is best.
    - List top move at each depth of search and whether it "matches" the move that was played (MM).
    - (Un)fair to count a "coincidence" if the played move matches at any depth?

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:
   - As the search progresses, they "change their minds" about which move is best.
   - List top move at each depth of search and whether it "matches" the move that was played (MM).
   - (Un)fair to count a "coincidence" if the played move matches at any depth?
   - Can set chess program to give values of multiple possible moves.

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:
   - As the search progresses, they "change their minds" about which move is best.
   - List top move at each depth of search and whether it "matches" the move that was played (MM).
   - (Un)fair to count a "coincidence" if the played move matches at any depth?
   - Can set chess program to give values of multiple possible moves.
3. Simple Principle: The wider the value gap between the best move and any other move, more likely a strong human player will find it.

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:
    - As the search progresses, they "change their minds" about which move is best.
    - List top move at each depth of search and whether it "matches" the move that was played (MM).
    - (Un)fair to count a "coincidence" if the played move matches at any depth?
    - Can set chess program to give values of multiple possible moves.
3. Simple Principle: The wider the value gap between the best move and any other move, more likely a strong human player will find it.
4. Judge a move "forced" if the value loss of the next-best move ("delta") is catastrophic.

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:
   - As the search progresses, they "change their minds" about which move is best.
   - List top move at each depth of search and whether it "matches" the move that was played (MM).
   - (Un)fair to count a "coincidence" if the played move matches at any depth?
   - Can set chess program to give values of multiple possible moves.
3. Simple Principle: The wider the value gap between the best move and any other move, more likely a strong human player will find it.
4. Judge a move "forced" if the value loss of the next-best move ("delta") is catastrophic.
5. Also cases where multiple moves have Equal-optimal Value (EV).

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:
   - As the search progresses, they "change their minds" about which move is best.
   - List top move at each depth of search and whether it "matches" the move that was played (MM).
   - (Un)fair to count a "coincidence" if the played move matches at any depth?
   - Can set chess program to give values of multiple possible moves.
3. Simple Principle: The wider the value gap between the best move and any other move, more likely a strong human player will find it.
4. Judge a move "forced" if the value loss of the next-best move ("delta") is catastrophic.
5. Also cases where multiple moves have Equal-optimal Value (EV).
6. Tally MM and EV cases and deltas over the games.

Some details on chess programs and model formation

1. (Try to) Reproduce the Accusation.
2. Observe Facts About Chess Programs:
    - As the search progresses, they "change their minds" about which move is best.
    - List top move at each depth of search and whether it "matches" the move that was played (MM).
    - (Un)fair to count a "coincidence" if the played move matches at any depth?
    - Can set chess program to give values of multiple possible moves.
3. Simple Principle: The wider the value gap between the best move and any other move, more likely a strong human player will find it.
4. Judge a move "forced" if the value loss of the next-best move ("delta") is catastrophic.
5. Also cases where multiple moves have Equal-optimal Value (EV).
6. Tally MM and EV cases and deltas over the games.
7. Find people to talk to...and compare results.

The Fortune and Blessings of Simplicity

The Fortune and Blessings of Simplicity

- How to build a *quantitative* model out of the simple principle?

The Fortune and Blessings of Simplicity

- How to build a *quantitative* model out of the simple principle?
- Simple elements **Strategy** and **Tactics** take us far.

The Fortune and Blessings of Simplicity

- How to build a *quantitative* model out of the simple principle?
- Simple elements **Strategy** and **Tactics** take us far.
- **Depth of Thinking** should be next.

The Fortune and Blessings of Simplicity

- How to build a *quantitative* model out of the simple principle?
- Simple elements **Strategy** and **Tactics** take us far.
- **Depth of Thinking** should be next.
- Do weaker players **prefer** weaker moves?

The Fortune and Blessings of Simplicity

- How to build a *quantitative* model out of the simple principle?
- Simple elements **Strategy** and **Tactics** take us far.
- **Depth of Thinking** should be next.
- Do weaker players **prefer** weaker moves?
- Or are they more easily **distracked?**

The Fortune and Blessings of Simplicity

- How to build a *quantitative* model out of the simple principle?
- Simple elements **Strategy** and **Tactics** take us far.
- **Depth of Thinking** should be next.
- Do weaker players **prefer** weaker moves?
- Or are they more easily **distracked?**
- How shall we handle the element of **Difficulty**?

The Fortune and Blessings of Simplicity

- How to build a *quantitative* model out of the simple principle?
- Simple elements **Strategy** and **Tactics** take us far.
- **Depth of Thinking** should be next.
- Do weaker players **prefer** weaker moves?
- Or are they more easily **distracked?**
- How shall we handle the element of **Difficulty**?
- **Recognition** "Versus" **Thinking**.

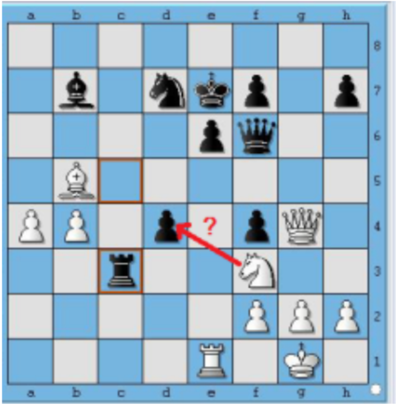
The Fortune and Blessings of Simplicity

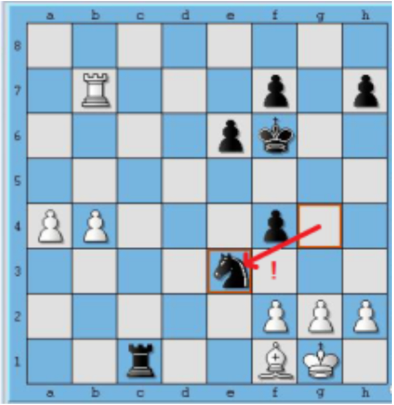
- How to build a *quantitative* model out of the simple principle?
- Simple elements **Strategy** and **Tactics** take us far.
- **Depth of Thinking** should be next.
- Do weaker players **prefer** weaker moves?
- Or are they more easily **distracked?**
- How shall we handle the element of **Difficulty**?
- **Recognition** "Versus" **Thinking**.
  - See the 2007 National Geographic documentary "My Brilliant Brain" with Susan Polgar (crux here).

The Fortune and Blessings of Simplicity

- How to build a *quantitative* model out of the simple principle?
- Simple elements **Strategy** and **Tactics** take us far.
- **Depth of Thinking** should be next.
- Do weaker players **prefer** weaker moves?
- Or are they more easily **distracked?**
- How shall we handle the element of **Difficulty**?
- **Recognition** "Versus" **Thinking**.
  - See the 2007 National Geographic documentary "My Brilliant Brain" with Susan Polgar (crux here).
  - We will try to glean comparable insight from numerical analytics.

Move Utilities Example (Kramnik-Anand, 2008)





Depths... Values by Stockfish 6

Move	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nd2	103	093	087	093	027	028	000	000	056	-007	039	028	037	020	014	017	000	006	000
Bxd7	048	034	-033	-033	-013	-042	-039	-050	-025	-010	001	000	-009	-027	-018	000	000	000	000
Qg8	114	114	-037	-037	-014	-014	-022	-068	-008	-056	-042	-004	-032	000	-014	-025	-045	-045	-050
...			...			...			...			...			...			...	
Nxd4	-056	-056	-113	-071	-071	-145	-020	-006	077	052	066	040	050	051	-181	-181	-181	-213	-213

Main Parameters and Inputs

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment.

Main Parameters and Inputs

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment. *Like Anatoly Karpov.*
- c for "consistency" in tactical minefields.

Main Parameters and Inputs

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment. *Like Anatoly Karpov.*
- c for "consistency" in tactical minefields. *Like Mikhail Tal.*

Main Parameters and Inputs

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment. *Like Anatoly Karpov.*
- c for "consistency" in tactical minefields. *Like Mikhail Tal.*
- h for "heave" or "Nudge"—obverse to depth of thinking.

Trained on all available in-person classical games in 2010–2019 with both players near the same Elo marker 1025, 1050, . . . , 2775, 2800, 2825.

Being retrained on new FIDE range **1400**. . . 2825, **from 1/1/25 on**.

Main Parameters and Inputs

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment. *Like Anatoly Karpov.*
- c for "consistency" in tactical minefields. *Like Mikhail Tal.*
- h for "heave" or "Nudge"—obverse to depth of thinking.

Trained on all available in-person classical games in 2010–2019 with both players near the same Elo marker 1025, 1050, . . . , 2775, 2800, 2825.

Being retrained on new FIDE range **1400**. . . 2825, **from 1/1/25 on**.

- Given an Elo rating R, "central slice" gives corresponding s_R, c_R, h_R.

Main Parameters and Inputs

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment. *Like Anatoly Karpov.*
- c for "consistency" in tactical minefields. *Like Mikhail Tal.*
- h for "heave" or "Nudge"—obverse to depth of thinking.

Trained on all available in-person classical games in 2010–2019 with both players near the same Elo marker 1025, 1050, . . . , 2775, 2800, 2825.

Being retrained on new FIDE range **1400**. . . 2825, **from 1/1/25 on**.

- Given an Elo rating R, "central slice" gives corresponding s_R, c_R, h_R.
- Only other input is the grid of move utilities $u_{i,d}$ at various depths d of search, further **scaled** to make (perceived) values v_i (and ρ_i).

Main Parameters and Inputs

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment. *Like Anatoly Karpov.*
- c for "consistency" in tactical minefields. *Like Mikhail Tal.*
- h for "heave" or "Nudge"—obverse to depth of thinking.

Trained on all available in-person classical games in 2010–2019 with both players near the same Elo marker 1025, 1050, . . . , 2775, 2800, 2825.

Being retrained on new FIDE range **1400**. . . 2825, **from 1/1/25 on**.

- Given an Elo rating R, "central slice" gives corresponding s_R, c_R, h_R.
- Only other input is the grid of move utilities $u_{i,d}$ at various depths d of search, further **scaled** to make (perceived) values v_i (and ρ_i).
- Then $\delta_i = v_1 - v_i$ is difference to best move.

Main Parameters and Inputs

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment. *Like Anatoly Karpov.*
- c for "consistency" in tactical minefields. *Like Mikhail Tal.*
- h for "heave" or "Nudge"—obverse to depth of thinking.

Trained on all available in-person classical games in 2010–2019 with both players near the same Elo marker 1025, 1050, ..., 2775, 2800, 2825.

Being retrained on new FIDE range **1400**...2825, **from 1/1/25 on**.

- Given an Elo rating R, "central slice" gives corresponding s_R, c_R, h_R.
- Only other input is the grid of move utilities $u_{i,d}$ at various depths d of search, further **scaled** to make (perceived) values v_i (and ρ_i).
- Then $\delta_i = v_1 - v_i$ is difference to best move.
- Other than these, **my model knows nothing about chess.**

One Wonky Slide: Log-Linear Versus Loglog-Linear

The generic **log-linear** model puts

$$\log\left(\frac{1}{p_i}\right) = \alpha + \beta u_i, \quad \text{or equivalently,} \quad \log\left(\frac{1}{p_i}\right) - \log\left(\frac{1}{p_1}\right) = \beta\delta_i$$

One Wonky Slide: Log-Linear Versus Loglog-Linear

The generic **log-linear** model puts

$$\log\left(\frac{1}{p_i}\right) = \alpha + \beta u_i, \quad \text{or equivalently,} \quad \log\left(\frac{1}{p_i}\right) - \log\left(\frac{1}{p_1}\right) = \beta\delta_i$$

- Solved by **softmax** giving $p_i = p_1 \cdot \exp(-\beta u_i)$.

One Wonky Slide: Log-Linear Versus Loglog-Linear

The generic **log-linear** model puts

$$\log\left(\frac{1}{p_i}\right) = \alpha + \beta u_i, \quad \text{or equivalently,} \quad \log\left(\frac{1}{p_i}\right) - \log\left(\frac{1}{p_1}\right) = \beta\delta_i$$

- Solved by **softmax** giving $p_i = p_1 \cdot \exp(-\beta u_i)$.
- Each p_i is represented as a **multiple** of the top probability p_1.

One Wonky Slide: Log-Linear Versus Loglog-Linear

The generic **log-linear** model puts

$$\log\left(\frac{1}{p_i}\right) = \alpha + \beta u_i, \quad \text{or equivalently,} \quad \log\left(\frac{1}{p_i}\right) - \log\left(\frac{1}{p_1}\right) = \beta\delta_i$$

- Solved by **softmax** giving $p_i = p_1 \cdot \exp(-\beta u_i)$.
- Each p_i is represented as a **multiple** of the top probability p_1.
- Ubiquitous in AI—but does not work for chess.

The **loglog-linear** model puts $\log\log(\frac{1}{p_i}) - \log\log(\frac{1}{p_1}) = \beta\delta_i$, i.e.:

$$\frac{\log(1/p_i)}{\log(1/p_1)} = \exp(\beta\delta_i).$$

One Wonky Slide: Log-Linear Versus Loglog-Linear

The generic **log-linear** model puts

$$\log\left(\frac{1}{p_i}\right) = \alpha + \beta u_i, \quad \text{or equivalently,} \quad \log\left(\frac{1}{p_i}\right) - \log\left(\frac{1}{p_1}\right) = \beta\delta_i$$

- Solved by **softmax** giving $p_i = p_1 \cdot \exp(-\beta u_i)$.
- Each p_i is represented as a **multiple** of the top probability p_1.
- Ubiquitous in AI—but does not work for chess.

The **loglog-linear** model puts $\log\log(\frac{1}{p_i}) - \log\log(\frac{1}{p_1}) = \beta\delta_i$, i.e.:

$$\frac{\log(1/p_i)}{\log(1/p_1)} = \exp(\beta\delta_i).$$

- Gives $p_i = p_1^{\exp(\beta\delta_i)}$.

One Wonky Slide: Log-Linear Versus Loglog-Linear

The generic **log-linear** model puts

$$\log\left(\frac{1}{p_i}\right) = \alpha + \beta u_i, \quad \text{or equivalently,} \quad \log\left(\frac{1}{p_i}\right) - \log\left(\frac{1}{p_1}\right) = \beta\delta_i$$

- Solved by **softmax** giving $p_i = p_1 \cdot \exp(-\beta u_i)$.
- Each p_i is represented as a **multiple** of the top probability p_1.
- Ubiquitous in AI—but does not work for chess.

The **loglog-linear** model puts $\log\log(\frac{1}{p_i}) - \log\log(\frac{1}{p_1}) = \beta\delta_i$, i.e.:

$$\frac{\log(1/p_i)}{\log(1/p_1)} = \exp(\beta\delta_i).$$

- Gives $p_i = p_1^{\exp(\beta\delta_i)}$.
- So p_i are represented as **powers** of the best-move probability p_1.

One Wonky Slide: Log-Linear Versus Loglog-Linear

The generic **log-linear** model puts

$$\log\left(\frac{1}{p_i}\right) = \alpha + \beta u_i, \quad \text{or equivalently,} \quad \log\left(\frac{1}{p_i}\right) - \log\left(\frac{1}{p_1}\right) = \beta\delta_i$$

- Solved by **softmax** giving $p_i = p_1 \cdot \exp(-\beta u_i)$.
- Each p_i is represented as a **multiple** of the top probability p_1.
- Ubiquitous in AI—but does not work for chess.

The **loglog-linear** model puts $\log\log(\frac{1}{p_i}) - \log\log(\frac{1}{p_1}) = \beta\delta_i$, i.e.:

$$\frac{\log(1/p_i)}{\log(1/p_1)} = \exp(\beta\delta_i).$$

- Gives $p_i = p_1^{\exp(\beta\delta_i)}$.
- So p_i are represented as **powers** of the best-move probability p_1.
- In place of $\beta\delta_i$, I really have $\left(\frac{\delta_i - h\rho_i}{s}\right)^c$, with h tightly clamped.

Karpov & Tal at Montreal "Tourney of Stars" 1979

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s = \mathbf{0.016}$, $c = \mathbf{0.307}$.

Karpov & Tal at Montreal “Tourney of Stars” 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s = \mathbf{0.016,}\ c = \mathbf{0.307}$.
- Tal (per Stockfish 11): $s = \mathbf{0.026,}\ c = \mathbf{0.365}$.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s = \mathbf{0.016}$**,** $c = \mathbf{0.307}$.
- Tal (per Stockfish 11): $s = \mathbf{0.026}$**,** $c = \mathbf{0.365}$.
- Lower s is better—so Karpov was more "Karpovian."

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s = \mathbf{0.016}$, $c = \mathbf{0.307}$.
- Tal (per Stockfish 11): $s = \mathbf{0.026}$, $c = \mathbf{0.365}$.
- Lower s is better—so Karpov was more "Karpovian."
- Higher c is better—so my model with Tal's parameters would make fewer large mistakes.

Are these grainy parameters enough to mimic human tendencies?

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s =$ **0.016**, $c =$ **0.307**.
- Tal (per Stockfish 11): $s =$ **0.026**, $c =$ **0.365**.
- Lower s is better—so Karpov was more "Karpovian."
- Higher c is better—so my model with Tal's parameters would make fewer large mistakes.

Are these grainy parameters enough to mimic human tendencies?

- IPRs: Karpov **2625 +- 155**, Tal **2730 +- 185**.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s =$ **0.016,** $c =$ **0.307**.
- Tal (per Stockfish 11): $s =$ **0.026,** $c =$ **0.365**.
- Lower s is better—so Karpov was more "Karpovian."
- Higher c is better—so my model with Tal's parameters would make fewer large mistakes.

Are these grainy parameters enough to mimic human tendencies?

- IPRs: Karpov **2625 +- 155**, Tal **2730 +- 185**.
- Whole tourney IPR is (only!) **2575 +- 50** ($s = 0.041$, $c = 0.385$).

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s =$ **0.016**, $c =$ **0.307**.
- Tal (per Stockfish 11): $s =$ **0.026**, $c =$ **0.365**.
- Lower s is better—so Karpov was more "Karpovian."
- Higher c is better—so my model with Tal's parameters would make fewer large mistakes.

Are these grainy parameters enough to mimic human tendencies?

- IPRs: Karpov **2625 +- 155**, Tal **2730 +- 185**.
- Whole tourney IPR is (only!) **2575 +- 50** ($s = 0.041$, $c = 0.385$).
- Average Elo of players, **2621**, is within error bars.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s = \mathbf{0.016}$, $c = \mathbf{0.307}$.
- Tal (per Stockfish 11): $s = \mathbf{0.026}$, $c = \mathbf{0.365}$.
- Lower s is better—so Karpov was more "Karpovian."
- Higher c is better—so my model with Tal's parameters would make fewer large mistakes.

Are these grainy parameters enough to mimic human tendencies?

- IPRs: Karpov **2625 +- 155**, Tal **2730 +- 185**.
- Whole tourney IPR is (only!) **2575 +- 50** ($s = 0.041$, $c = 0.385$).
- Average Elo of players, **2621**, is within error bars. Surprise is that the IPR is not near 2700s range.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s = \mathbf{0.016}$**,** $c = \mathbf{0.307}$.
- Tal (per Stockfish 11): $s = \mathbf{0.026}$**,** $c = \mathbf{0.365}$.
- Lower s is better—so Karpov was more "Karpovian."
- Higher c is better—so my model with Tal's parameters would make fewer large mistakes.

Are these grainy parameters enough to mimic human tendencies?

- IPRs: Karpov **2625 +- 155**, Tal **2730 +- 185**.
- Whole tourney IPR is (only!) **2575 +- 50** ($s = 0.041$, $c = 0.385$).
- Average Elo of players, **2621**, is within error bars. Surprise is that the IPR is not near 2700s range. Today's elite regularly hit 2800+.

Simplicity and Public Outreach

Simplicity and Public Outreach

- Originally I intended to use *distributional distance measures*, of which **fidelity** is one.

Simplicity and Public Outreach

- Originally I intended to use *distributional distance measures*, of which **fidelity** is one.
- I realized that results would be difficult to **explain**.

Simplicity and Public Outreach

- Originally I intended to use *distributional distance measures*, of which **fidelity** is one.
- I realized that results would be difficult to **explain**.
- Hence I mapped everything to rolling dice, with math known since 1800.

Simplicity and Public Outreach

- Originally I intended to use *distributional distance measures*, of which **fidelity** is one.
- I realized that results would be difficult to **explain**.
- Hence I mapped everything to rolling dice, with math known since 1800.
- Other simple matters: Should Metrics Be Linear?

Simplicity and Public Outreach

- Originally I intended to use *distributional distance measures*, of which **fidelity** is one.
- I realized that results would be difficult to **explain**.
- Hence I mapped everything to rolling dice, with math known since 1800.
- Other simple matters: Should Metrics Be Linear?
- "Pandemic Lag" in updating ratings.

Simplicity and Public Outreach

- Originally I intended to use *distributional distance measures*, of which **fidelity** is one.
- I realized that results would be difficult to **explain**.
- Hence I mapped everything to rolling dice, with math known since 1800.
- Other simple matters: Should Metrics Be Linear?
- "Pandemic Lag" in updating ratings.
- Are their players who are majorly better, relative to their peers at slow chess, in fast chess?

Simplicity and Public Outreach

- Originally I intended to use *distributional distance measures*, of which **fidelity** is one.
- I realized that results would be difficult to **explain**.
- Hence I mapped everything to rolling dice, with math known since 1800.
- Other simple matters: Should Metrics Be Linear?
- "Pandemic Lag" in updating ratings.
- Are their players who are majorly better, relative to their peers at slow chess, in fast chess?
- [Segue to public-outreach cases and demos]

Simplicity and Public Outreach

- Originally I intended to use *distributional distance measures*, of which **fidelity** is one.
- I realized that results would be difficult to **explain**.
- Hence I mapped everything to rolling dice, with math known since 1800.
- Other simple matters: Should Metrics Be Linear?
- "Pandemic Lag" in updating ratings.
- Are their players who are majorly better, relative to their peers at slow chess, in fast chess?
- [Segue to public-outreach cases and demos]
- Q & A — And Thanks.